PATENT
1509-477

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: Takahiko KAWATANI

Serial No.: NEW

Filed: March 4, 2004

For: DOCUMENT AND PATTERN CLUSTERING METHOD AND APPARATUS

CLAIM FOR PRIORITY AND SUBMISSION OF PRIORITY DOCUMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

March 4, 2004

Sir:

In accordance with the provisions of 35 U.S.C. §119, Applicant hereby claims the benefit of the filing dates of Japanese Patent Application Nos. 2003-105867, filed March 5, 2003, and 2004-30629, filed February 6, 2004. A certified copy of JP 2003-105867 is attached; a certified copy of JP 2004-30629 will be filed in due course. The Examiner is respectfully requested to acknowledge Applicant's claim for priority, as well as receipt of the certified copy of the priority document.

Respectfully submitted,
LOWE HAUPTMAN GILMAN & BERNER, LLP

By: Allan M. Lowe
Allan M. Lowe, Reg. No. 19,641

1700 Diagonal Road, Suite 300
Alexandria, VA 22314
703-684-1111 telephone
703-518-5499 telecopier
AML:rk

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　３月　５日

出願番号
Application Number: 特願２００３－１０５８６７

[ST. 10/C]: [ＪＰ２００３－１０５８６７]

出願人
Applicant(s): ヒューレット・パッカード・カンパニー

２００３年　７月２３日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３０５８６２０

【書類名】　特許願
【提出日】　平成15年 3月 5日
【整理番号】　200309904
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 17/21
【発明者】
　【住所又は居所】　東京都杉並区高井戸東３丁目２９番２１号　日本ヒューレット・パッカード株式会社内
　【氏名】　川谷　隆彦
【特許出願人】
　【識別番号】　398038580
　【氏名又は名称】　ヒューレット・パッカード・カンパニー
【代理人】
　【識別番号】　100082946
　【弁理士】
　【氏名又は名称】　大西　昭広
【手数料の表示】
　【予納台帳番号】　061492
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　0300115
【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　文書のクラスタリング方法及び装置

【特許請求の範囲】

【請求項１】

以下の（ａ）から（ｅ）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、文書間の関係に基づき文書をグループ化するクラスタリング方法、

（ａ）前記文書集合と共通性の高い文書を前記文書集合から選択しクラスターとするステップ、

（ｂ）前記選択された文書と類似性の高い全ての文書を前記クラスターに追加するステップ、

（ｃ）前記クラスターに存在する文書と共通性の高い全ての文書を前記クラスターに繰返し追加するステップ、

（ｄ）既に求められたクラスターと共通性の高い文書を除外した文書集合を前記文書集合として前記（ａ）、（ｂ）及び（ｃ）のステップを繰返すステップ、

（ｅ）前記クラスターが重複する場合に、冗長なクラスターを除去するステップ。

【請求項２】

前記（ａ）のステップは、さらに以下の（ａ－１）から（ａ－４）のステップを含むことを特徴とする請求項１に記載の方法、

（ａ－１）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を１、他の値は０とする文書セグメントベクトルを生成するステップと、

（ａ－２）文書セグメントベクトルより生成する共起行列を前記文書集合の各文書に対して求めるステップと、

（ａ－３）前記文書集合に含まれる全文書から文書頻度行列を求めるステップと、

（ａ－４）前記共起行列及び文書頻度行列を用いて、前記文書集合と共通性の高い文書を選択するステップ。

【請求項３】

前記（ｃ）のステップは、さらに以下の（ｃ－１）から（ｃ－３）のステップを含むことを特徴とする請求項１に記載の方法、

（ｃ－１）各用語の入力文書集合における出現文書数と前記クラスター内での出現文書数から文書頻度比を求めるステップ、

（ｃ－２）前記クラスター内での出現文書数や前記文書頻度比の情報を用いて用語や用語対を選択し、選択された用語や用語対を用いて文書共通度を求めるステップ、

（ｃ－３）前記文書共通度に基づいて、前記前記選択された文書と共通性の高い文書を選択するステップ。

【請求項４】

前記（ｄ）のステップは、さらに以下の（ｄ－１）から（ｄ－４）のステップを含むことを特徴とする請求項１に記載の方法、

（ｄ－１）各文書と各クラスターとの文書共通度を求めるステップ、

（ｄ－２）各クラスター毎に、前記文書共通度が、着目クラスターに対しては所定値よりも大きく、他の全てのクラスターに対して所定値よりも小さい文書数を求めるステップ、

（ｄ－３）前記文書数が所定値よりも小さいクラスターが存在する場合には、最少の前期文書数を有するクラスターを削除するステップ、

（ｄ－４）前記（ｄ－１）から（ｄ－３）を繰返すステップ。

【請求項５】

以下の（ａ）から（１）のステップを有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、同じ話題を有する文書をグループ化するクラスタリング方法、

（ａ）前記文書セグメント毎に、前記文書セグメントに出現する用語に対応する成分の値を１、他の値は０とする文書セグメントベクトルを生成するステップと、

（ｂ）文書セグメントベクトルより生成する共起行列を前記文書集合の各文書に対して求めるステップと、

（ｃ）前記文書集合に含まれる全文書から文書頻度行列を求めるステップと、

（ｄ）その時点で求められているどのクラスターに対しても文書共通度が一定値以下の文書の集合に対して共通共起行列を求めるステップと、

（ｅ）前記その時点で求められているどのクラスターに対しても文書共通度が一定値以下の文書の集合の中から（ｄ）で求められた共通共起行列を用いてクラスターの種となる文書を抽出し、クラスターの種文書の近隣文書からなる初期クラスターを作成するステップと、

（ｆ）その時点で着目クラスターに一時的に帰属する文書集合から共通共起行列と文書頻度行列を求めるステップと、

（ｇ）（ｃ）及び（ｆ）で求められる文書頻度行列を比較して、各用語及び用語対の着目クラスターに対する特有さの情報を求めるステップと、

（ｈ）各文書の着目クラスターに対する文書共通度を、（ｆ）で求められる共通共起行列と、（ｇ）で求められる特有さの情報から求めた各用語及び用語対の重みとを用いて求め、一定値以上の文書共通度を有する文書を着目クラスターに一時的に帰属させるステップと、

（ｉ）上記（ｆ）から（ｈ）までのステップを着目クラスターに一時的に帰属する文書数が増えなくなるまで繰返すステップと、

（ｊ）上記（ｄ）から（ｉ）までのステップを、どのクラスターに対しても文書共通度が一定値以下の文書の数が０になるか、もしくはその数が一定値以下で前回の繰り返しの時と同じになるまで繰返すステップと、

（ｋ）各文書の各クラスターに対する文書共通度をもとに、各文書の帰属するクラスターを決定するステップと、

（ｌ）冗長なクラスターの有無をチェックし、冗長なクラスターが存在すれば除去したうえで、各文書の帰属するクラスターを改めて決定するステップ。

【請求項６】

前記出現する用語の種類数がＭで与えられ、Ｒ個の文書からなる文書集合Ｄにおいて、ｒ番目の文書をD_r、D_rの文書セグメント数をY_r、D_rのｙ番目の文書セグメントベクトルをd_{ry}（d_{ry1}，．．，d_{ryM}）^{T}とすると、ここで、Ｔはベクトルの転置を表す、文書D_rの前記共起行列S_rは、

【数１】

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}{}^T \quad \cdots (1)$$

で与えられることを特徴とする請求項５に記載の方法。

【請求項７】

文書集合Ｄの文書頻度行列の各成分は、文書集合Ｄ中の各文書の共起行列の対応する成分がゼロでない文書数であることを特徴とする請求項６に記載の方法。

【請求項８】

文書集合Ｄの共通共起行列は、ｍｎ成分が以下のように決定される行列Ｔ

【数２】

$$T_{mn} = \prod_{\substack{r=1 \\ S^r{}_{mn}>0}}^{R} S^r{}_{mn} \quad \cdots (2)$$

をもとに、ｍｎ成分が

$T^A{}_{mn} = T_{mn}$,　　$U_{mn} \geq A$の時

$T^A{}_{mn} = 0$　　それ以外のとき

によって決定される行列T^Aにより、もしくはｍｎ成分が

$Q^A{}_{mn} = log(T^A{}_{mn})$　　$T^A{}_{mn} > 1$の時

$Q^A{}_{mn} = 0$　　それ以外のとき

によって決定される行列Q^Aにより与えられることを特徴とする請求項６、７に記載の方法。

【請求項９】

共起行列をS^Pとする文書Ｐの文書集合Ｄに対する文書共通度は、z_{mm}、z_{mn}をそれぞれ用語ｍ、用語対ｍ、ｎに対する重みとして

【数３】

$$com_I(D,P;Q^A) = \frac{\sum_{m=1}^{M} z_{mm} Q^A{}_{mm} S^P{}_{mm}}{\sqrt{\sum_{m=1}^{M} z_{mm} (Q^A{}_{mm})^2} \sqrt{\sum_{m=1}^{M} z_{mm} (S^P{}_{mn})^2}} \quad \cdots (3)$$

もしくは、

【数４】

$$com_q(D,P;Q^A)=\frac{\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}Q^A{}_{mn}S^P{}_{mn}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}(Q^A{}_{mn})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}(S^P{}_{mn})^2}}$$

．．．（４）

もしくは数式（３）、数式（４）において行列ＱＡの代わりに行列ＴＡを用いた式により与えられることを特徴とする請求項６、７、８に記載の方法。

【請求項１０】

クラスターの種となる文書の抽出と初期クラスターの作成は以下の（ａ）から（ｄ）のステップを経て行われることを特徴とする請求項６、７、８、９に記載の方法。

（ａ）その時点で求められているどのクラスターに対しても文書共通度が一定値以下の文書の集合から求められる共通共起行列を用いて、前記文書集合中の各文書の文書集合共通度、もしくは情報共通量を求めるステップと、

（ｂ）（ａ）で求められた文書集合共通度、もしくは情報共通量の大きい一定個の文書をクラスターの種の候補として抽出するステップと、

（ｃ）クラスターの種の各候補について、前記文書集合中の各文書との類似度を求め、類似度が一定値以上となる文書を近隣文書として求めるステップと、

（ｄ）クラスターの種の各候補の中から近隣文書の最も多い候補を選択してクラスターの種の文書とし、その近隣文書を初期クラスターとするステップ。

【請求項１１】

共起行列をＳＰとする文書Ｐの文書集合Ｄに対する情報共通量は

【数５】

$$comInfo(D,P)=\sum_{m=1}^{M}\sum_{n=1}^{M}z_{mn}Q^A{}_{mn}S^P{}_{mn}$$

．．．（５）

もしくは数式（５）において行列ＱＡの代わりに行列ＴＡを用いた式により与えられることを特徴とする請求項１０に記載の方法。

【請求項１２】

各用語及び用語対の着目クラスターに対する特有さの情報と重みの決定は、以下の（ａ）から（ｄ）のステップを経て行われることを特徴とする請求項６、７、８、９、１０、１１に記載の方法。

（ａ）入力文書全体から求められる文書頻度行列の各成分の、その時点で着目クラスターに一時的に帰属する文書集合から求められる文書頻度行列の対応する成分に対する比を、対角成分の場合には用語文書頻度比として、非対角成分の場合には用語対文書頻度比として求めるステップと、

（ｂ）その時点で着目クラスターに一時的に帰属する文書集合において、最も頻度の高い一定個の用語の中で用語文書頻度比が小さい一定個の用語を選択し、それらの用語文書頻度比の平均を平均文書頻度比として求める、もしくは、最も頻度の高い一定個の用語や用語対の中で用語もしくは用語対文書頻度比が小さい一定個の用語や用語対を選択し、それらの用語もしくは用語対文書頻度比の平均を平均文書頻度比として求めるステップと、

（ｃ）平均文書頻度比を各用語または用語対の用語文書頻度比または用語対文書頻度比で除した値を各用語または用語対の特有さの情報として求めるステップと

（ｄ）特有さの情報を変数とする関数によって用語または用語対の重みを決定するステップ。

【請求項１３】

文書入力部、文書前処理部、文書情報処理部、及び出力部を有する装置において、一つまたは複数の文書セグメントを持つ複数の文書から成る入力文書集合に対して、文書間の関係に基づき文書をグループ化するために、

（ａ）前記文書集合と共通性の高い文書を前記文書集合から選択しクラスターとするクラスター選択部、

（ｂ）前記選択された文書と類似性の高い全ての文書を前記クラスターに追加する追加部、

（ｃ）前記クラスターに存在する文書と共通性の高い全ての文書を前記クラスターに繰返し追加するクラスター成長部、

（ｄ）既に求められたクラスターと共通性の高い文書を除外した文書集合を前記

文書集合として前記（ａ）、（ｂ）及び（ｃ）のステップを繰返す繰返し部、
（ｅ）前記クラスターが重複する場合に、冗長なクラスターを除去する除去部、
を動作させる為のプログラム。

【請求項１４】

文書入力部、文書前処理部、文書情報処理部、及び出力部を有する装置において、以下の（ａ）から（ｄ）の手段を有する、一つまたは複数の文書セグメントを持つ複数の文書から成る文書集合に対して、文書間の関係に基づき文書をグループ化するクラスタリング装置、
（ａ）前記文書集合と共通性の高い文書を前記文書集合から選択しクラスターとする手段、
（ｂ）前記選択された文書と類似性の高い全ての文書を前記クラスターに追加する手段、
（ｃ）前記クラスターに存在する文書と共通性の高い全ての文書を前記クラスターに繰返し追加する手段、
（ｄ）既に求められたクラスターと共通性の高い文書を除外した文書集合を前記文書集合として前記（ａ）、（ｂ）及び（ｃ）のステップを繰返す手段、
（ｅ）前記クラスターが重複する場合に、冗長なクラスターを除去する手段。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】

本発明は文書のクラスタリングをはじめとする自然言語処理に関するものであり、前記処理の高性能化を図ることによって文書からの情報の抽出を容易にするものである。

【０００２】

【従来の技術】

文書クラスタリングは入力された文書群を文書の内容によって幾つかのグループに分割する技術である。クラスタリング技術は古くから研究されてきており、これまでに考案された方法についてはＣ．Ｄ．ＭａｎｎｉｎｇとＨ．Ｓｃｈｕｔｚｅによって著されたＦｏｕｎｄａｔｉｏｎｓ　ｏｆ　Ｓｔａｔｉｓｔｉｃａｌ

Ｎａｔｕｒａｌ　Ｌａｎｇｕａｇｅ　Ｐｒｏｃｅｓｓｉｎｇ（Ｔｈｅ　ＭＩＴ　Ｐｒｅｓｓ、１９９９年）に体系的に紹介されている。先ず、クラスタリングは、各文書が各クラスターに帰属する確率を求めるソフトクラスタリング、各クラスターに帰属するか否かを求めるハードクラスタリングに大別される。後者については、さらに、階層的な手法と非階層的な手法とに分類される。階層的な手法は、さらにボトムアップのアプローチとトップダウンのアプローチに分けられる。前者では、初期状態として各文書をクラスターの核とし、最も近いクラスター同士をマージするという処理を繰り返す。これにより文書集合は木構造で表現されるようになる。クラスター間の近さの尺度、即ち類似度を図る方法として単一リンク法、完全リンク法、グループ平均法が知られている。これらは何れも２文書間の類似度をもとに算出されるものである。後者では、全文書が１つのクラスターに属するという状態から出発し、例えばひとつのクラスター中のあらゆる文書対の中で最も低い類似度が閾値以下の場合、そのクラスターを分割するという処理を繰り返す。

【０００３】

非階層的な手法では、予め指定された数のクラスターが何らかの基準を満たすように作成される。よく知られている方法の例を以下に示す。

ステップ１：指定されたクラスター数の文書をランダムに選択して各クラスターの中心とするステップ、

ステップ２：各文書について各クラスター中心との近さを求め、各文書を最も近いクラスターに帰属させるステップ、

ステップ３：各クラスターに帰属する文書ベクトルの平均により各クラスターの中心を求めるステップ、

ステップ４：ステップ２）の処理を実行し、各文書の帰属するクラスターに変化がなければ終了、そうでなければステップ３へいくステップ、からなる方法である。

【０００４】

【発明が解消しようとする課題】

従来の文書クラスタリング技術は２つの大きな問題を抱えている。ひとつは求め

られるクラスターの数の問題である。文書クラスタリングでは求められるクラスターの数は入力された文書集合中の文書が述べている話題の数と同じでなければならない。前述のようにボトムアップの階層的なクラスタリング処理では、各クラスターは一つの文書から成る状態から出発し、最も近いクラスター同士をマージする処理を繰り返して最後は全文書が一つのクラスターに属することになる。従って、話題の数と同じ数のクラスターを得るにはクラスターのマージを打ち切ることが必要となる。これは、クラスターのマージ処理において類似度が閾値以下のクラスター対についてはマージを行わないようにすることにより実現可能である。しかしながら、実際には上述の閾値を如何に決めるかが難しい問題となっている。閾値が不適切であれば正しい数のクラスターは得られない。同様にトップダウンのクラスタリング処理では、ひとつのクラスター中のあらゆる文書対の中で最も低い類似度が閾値以上の場合にはクラスターの分割は行わないようにすれば、原理的には話題の数と同じ数のクラスターが得られる筈である。

【０００５】

しかしながら、この場合にも閾値をどのように決めるかが難しい問題となっている。また、非階層的なクラスタリングでは、ユーザーは与えられた文書集合を何個のクラスターに分割するかの情報を事前に入力することを求められる。しかしながら、クラスター数の情報は入力文書集合の事前の知識なしには正確に与えることは不可能である。このように入力文書集合から正しい数のクラスターを得ることは難しい問題となっている。非階層的なクラスタリングでクラスター数を正しく推測する試みもなされており、性能は向上しているが完璧ではない（X. Liu, Y. Gong, W. Xu and S. Zhu. Document Clustering with Cluster Refinement and Model Selection Capabilities. In Proceedings of the 25th Annual International ACM－SIGIR Conference on Research and Development in Information Retrieval, pp. 191－198. Tampere, Finland, August, ２００２.）。

【０００６】

２番目の問題はクラスタリングにおける精度の問題である。これは、ひとつのクラスターに属する文書が同じ話題を述べているかどうかの問題である。クラスタリング処理では、通常は文書は、各成分が文書中に現れる各用語の有無、もしくは出現頻度で与えられるベクトルで表現され、２つのクラスター間の類似度は、異なるクラスターに属する文書対の余弦類似度をもとに、ある文書とクラスター間の類似度はその文書のベクトルとそのクラスターに属する文書の平均ベクトルとの間の距離（例えばユークリッド距離）をもとに求められていた。従来のクラスタリング処理では、上記の余弦類似度やユークリッド距離を求めるときに、どのような用語がそのクラスターに重要なのかを検証することなく、各文書で得られたベクトルをそのまま用いることが多かった。そのため、各クラスターに本質的でない情報を表す用語や用語対の存在がクラスタリングの精度に影響を及ぼしえた。

【０００７】

【課題を解決するための手段】

ところで、文書クラスタリングは各文書が記述する話題によって文書をグループ化するものであるから、ひとつのクラスターに属する文書（クラスター文書集合と呼ぶ）は同じ話題について述べている筈である。従って、クラスター文書集合は何らかの共通性を有する筈である。また、各話題にはその話題には多く出現し、他の話題にはあまり出現しない話題特有の用語や用語対が存在する筈である。従って、各クラスターには用語や用語対の出現傾向に違いが存在する筈である。このようなことから、本発明ではクラスタリングの精度を高めるために、クラスタリングの過程で、

Ａ）：着目クラスター文書集合の共通情報を抽出し、この共通情報との近さにより各文書の着目クラスターに対する近さ（文書共通度）を求める。

Ｂ）：着目クラスターに特有でない用語や用語対を検出し、上記文書共通度の算出においてはそれらの影響を排除する。

という手段を導入する。

【０００８】

従来の階層的な処理では、クラスターのマージや分割が頻繁に繰り返される。また、従来の非階層的な処理では、クラスターのメンバーがやはり頻繁に入れ替わる。このような状況では各クラスターの共通情報、クラスターに非特有な用語や用語対の検出には無理がある。そこで、本発明ではクラスタリングの全体の手順として以下を採用する。

【０００９】

ステップ１：最初の繰り返しにおいては全文書から、２回目以降の繰り返しにおいては、その時点で存在するどのクラスターに対しても文書共通度が一定値以下の文書の中から、クラスターの種の候補となる文書を検出する。

ステップ２：先ず、クラスターの種の各候補文書について、全文書との類似度を求め一定値以上の類似度を有する文書を抽出する。その文書数が最も大きくなる候補文書をクラスターの種とし、その文書集合によりクラスターを形成する。

ステップ３：その時点でのクラスター文書集合と各文書との間で文書共通度を求め、一定値以上の文書共通度を有する文書をそのクラスターに一時的に帰属させることによりクラスターを成長させる。クラスターに一時的に属する文書数が一定になればステップ４へ移行する。そうでなければ本ステップを繰返す。

ステップ４：終了条件を満たせばステップ５へ移行する。そうでなければステップ１に戻って続行する。

ステップ５：各文書について各クラスターへの文書共通度を求め、文書共通度の最も高くなるクラスターに帰属させる。

ステップ６：１つの話題に２つ以上のクラスターが対応していないかどうかを検出する。そのようなクラスターがあれば冗長なクラスターとして削除し、各文書の帰属するクラスターを求めなおす。

【００１０】

上記のクラスタリング手順において、前記Ａ）の共通情報を用いる文書共通度の算出、及びＢ）の着目クラスターに特有でない用語や用語対の検出はステップ３と５において行われることになる。Ａ）については、その時点での着目クラスターに一時的に帰属している文書から共通情報を抽出することになる。共通情報の抽出と利用は特願２００２－３２６１５７で述べられている方法を援用すること

ができる。基本的な考え方は次の通りである。いま、着目クラスターがＲ個の文書から成るものとし、各文書から一つづつ文を取り出してＲ個の文からなる文の組を作ったとする。このような文の組は各文書の文の数の積通り存在することになる。ここでは、着目する文の組において、Ｒ個の文のうちのＡ個の文に現れる用語を共通用語、共通用語で構成された文を共通文と呼ぶこととする。ここで、全ての文の組で共通文を作り、共通文の集合を構成したとする。このような共通文の集合は着目クラスターの共通の話題の内容を示すものと考えられる。従って、各文書と共通文集合との間で何らかの手段で類似度を求めることができれば、それは各文書の着目クラスターの共通話題への近さを表わすことになる。

【００１１】

またＢ）の着目クラスターに特有でない用語や用語対の検出は次のような考え方で行うものである。種文書が話題ｉの着目クラスターの成長の過程を考える。話題ｉを述べている文書数は、文書集合全体にはｃ０個、着目クラスターの文書集合にはｃ個存在したとする。また、用語ｍを含む文書数は、文書集合全体ではＵ０ｍｍ個、着目クラスターの文書集合ではＵｍｍ個存在したとする。用語ｍが話題ｉの特有用語の時には用語ｍは話題ｉの文書のみに現れるので、

【００１２】

【数６】

$$U^0_{mm}/U_{mm} \approx c_0/c$$

【００１３】

となり、非特有の時には話題ｉ以外の話題の文書にも現れるので、

【００１４】

【数７】

$$U^0_{mm}/U_{mm} > c_0/c$$

【００１５】

となる筈である。従って、ｃ０／ｃを適当な方法で求めることができれば用語ｍが話題ｉに特有か否かを判断することができる。Ｕ０ｍｍ／Ｕｍｍを用語ｍの文

書頻度比と呼ぶことにすると、本発明では、着目クラスターの文書集合において最も頻度の高い一定個の用語のうち、文書頻度比の値の小さな一定個は話題ｉの特有用語とみなし、これらの用語の文書頻度比の平均ｃ’をｃ０／ｃの推測値とした。結局、αをパラメータとして

【００１６】

【数８】

$$U^0_{mm}/U_{mm} > \alpha c'$$

【００１７】

を満たす用語ｍは話題ｉには特有な用語ではないと判断できる。

同様に、用語ｍ、ｎを含む文書数は、文書集合全体ではＵ０ｍｎ個、着目クラスターの文書集合ではＵｍｎ個存在したとして、

【００１８】

【数９】

$$U^0_{mn}/U_{mn} > \alpha c'$$

【００１９】

を満たす用語対ｍ、ｎは話題ｉには特有な用語対ではないと判断できる。

文書共通度は着目クラスターに本質的でない用語や用語対の影響を受けにくくするためには、話題ｉには特有ではないと判断された用語、用語対は、各文書と着目クラスターの文書集合との文書共通度の算出に用いないようにすればよい。もしくは、

【００２０】

【数１０】

$$c'/(U^0_{mm}/U_{mm})$$

【００２１】

【数１１】

$$c'/(U^0_{mn}/U_{mn})$$

【００２２】

をそれぞれ用語ｍ、用語対ｍ、ｎの重みとして用いて文書共通度の算出を行ってもよい。このようにすることにより、話題ｉを述べた文書に対して文書共通度は

大きな値をとるようになる。その結果クラスタリングの精度の向上が期待できる。

【００２３】

上記のクラスタリングの全体手順においては、先ずクラスターの種となる文書をひとつ取り出し、ついでその種と同じ話題を記述する文書を検出して種を成長させるという処理を繰り返し行うこととなる。従って、種の文書の数が入力文書における話題の数と過不足なく一致すれば正しい数のクラスターが得られることになる。たとえステップ１において同じ話題に対して２つの種文書が検出されたにしても、ステップ６で冗長なクラスターを検出して除去するので正しい数のクラスターが得られる。また、ステップ１においてある話題に対して種文書が検出されない時にはクラスターの数が不足することになる。このような事態は、クラスタリングの精度が低く、検出されるべき話題の文書がすでに存在するクラスターとの文書類似度が高くなってしまった時に起きる。言い換えれば、このような事態は、クラスタリングの精度が低いために、ひとつのクラスターに本来有すべきでない話題を有する文書が混入することによって引き起こされる。しかしながら、本発明では上記Ａ）、Ｂ）の手段によってクラスタリングの精度を高めているので異なる話題の文書の混入の可能性は低く、求められるクラスターの数が少なくなるという事態は起こりにくい。本発明では正しい数のクラスターが得られる公算が非常に大きい。

【００２４】

【実施例】

図１は、本発明の概要を示すブロック図である。１１０は文書入力ブロック、１２０は文書前処理ブロック、１３０は文書情報処理ブロック、１４０は出力ブロックを示す。文書入力ブロック１１０には、処理したい文書集合が入力される。文書前処理ブロック１２０では、入力された文書の用語検出、形態素解析、文書セグメント区分け等が行われる。文書セグメントについて説明する。文書セグメントは文書を構成する要素であり、その最も基本的な単位は文である。英文の場合、文はピリオドで終わり、その後ろにスペースが続くので文の切出しは容易に行うことができる。その他の文書セグメントへの区分け法としては、ひとつの文

が複文からなる場合主節と従属節に分けておく方法、用語の数がほぼ同じになるように複数の文をまとめて文書セグメントとする方法、文書の先頭から含まれる用語の数が同じになるように文とは関係なく区分けする方法などがある。文書情報処理ブロック１３０は以下に詳細に説明するが、情報処理を行い、種の文書の検出、全文書と着目クラスターとの文書集合共通度の算出、各クラスターの特有でない用語、用語対の検出などクラスタリングに直接関わる処理を行う。出力ブロック１４０は文書情報処理ブロック１３０で得られた結果を、ディスプレー等の出力装置に出力する。

【００２５】

図２は与えられた文書集合に対して、クラスタリングを行う本発明の実施例を示す。この発明の方法は、汎用コンピュータ上でこの発明を組み込んだプログラムを走らせることによって実施することができる。図２は、そのようなプログラムを走らせている状態でのコンピュータのフローチャートである。ブロック２１は文書集合入力、ブロック２２は文書前処理、ブロック２３は全体文書情報抽出処理、ブロック２４はクラスターの種文書及び初期クラスターの決定、ブロック２５はクラスターの成長処理、ブロック２６は残存文書検出、ブロック２７は終了条件チェック、ブロック２８は残存文書の文書情報抽出処理、ブロック２９は帰属クラスター決定、ブロック３０は冗長クラスターの検出・除去、である。以下、英文文書を例に実施例を説明する。

【００２６】

先ず、文書集合入力２１において対象となる文書集合が入力される。文書前処理２２においては各入力文書に対して、用語検出、形態素解析、文書セグメント区分け、文書セグメントベクトル作成などの前処理が行われる。用語検出としては、各入力文書から単語、数式、記号系列などを検出する。ここでは、単語や記号系列などを総称して全て用語と呼ぶ。英文の場合、用語同士を分けて書く正書法が確立しているので用語の検出は容易である。次に、形態素解析では、各入力文書に対して用語の品詞付けなどの形態素解析を行う。文書セグメント区分けにおいて各入力文書に対して文書セグメントへの区分けを行う。文書セグメントベクトル作成では、先ず文書全体に出現する用語から作成すべきベクトルの次元数

および各次元と各用語との対応を決定する。この際に出現する全ての用語の種類にベクトルの成分を対応させなければならないということはなく、品詞付け処理の結果を用い、例えば名詞と動詞と判定された用語のみを用いてベクトルを作成するようにしてもよい。次いで、各文書セグメントに出現する用語に対応する成分のみが値１、他は０となるような文書セグメントベクトルを作成する。

【００２７】

全体文書情報抽出処理２３では、後段のクラスタリング処理に用いるデータを各文書と入力文書集合全体から求める。求めるデータは各文書の共起行列、前記共通文集合の共起行列（共通共起行列）、入力文書集合全体の文書頻度行列である。各文書の共起行列は用語の出現頻度、用語間の共起頻度を反映する行列である。文を文書セグメントとした場合について説明を続ける。ここでは、現れる用語集合が｛ｗ１，．．，ｗＭ｝で与えられ、Ｒ個の文書から成る入力文書集合Ｄを考える。さらに、ｒ番目の文書をＤｒとすると、ＤｒはＹｒ個の文からなるものとし、ｙ番目の文及びその文ベクトルをＤｒｙ、ｄｒｙ＝（ｄｒｙ１，．．，ｄｒｙＭ）Ｔとする。ここで、Ｔはベクトルの転置を表す。ｄｒｙは２値ベクトルであり、ｄｒｙｍはｍ番目の用語の有無を表す。文書のＤｒの共起行列をＳｒとすると、これは

【００２８】

【数１２】

$$S^r = \sum_{y=1}^{Y_r} d_{ry} d_{ry}^T \quad \cdots (1)$$

【００２９】

$S^r{}_{mn} = \sum_{y=1}^{Y_r} d_{rym} d_{ryn}$ で与えられる。式（１）から分かるように、Ｓｒのmn成分はに

【００３０】

【数１３】

$$T_{mn} = \prod_{r=1}^{R}{}_{S^r{}_{mn}>0} S^r{}_{mn}$$

【００３１】

さらに、各用語、用語共起の文書頻度を格納した行列Ｕ０を求める。Ｕ０ｍｍ、Ｕ０ｍｍはそれぞれ用語ｍの出現する文書数、用語ｍ、ｎの共起する文書数となる。このような行列Ｔ、Ｕ０を用いて共通共起行列ＴＡを求める。共通共起行列ＴＡのｍｎ成分は以下のように決められる。

$T^A_{mn} = T_{mn}$,　　$U^0_{mn} \geq A$のとき

$T^A_{mn} = 0$　　それ以外

Ａはパラメータであり、実験的に決められる。

【００３２】

また、ｍｎ成分が以下のように与えられる行列ＱＡを

$Q^A_{mn} = \log(T^A_{mn})$　　$T^A_{mn} > 1$のとき

$Q^A_{mn} = 0$　　それ以外

により定義し、共通共起行列として用いてもよい。

【００３３】

クラスターの種文書及び初期クラスターの決定２４では、前記ステップ１及び２に対応する処理を行う。ここで、その時点で存在するどのクラスターに対しても文書共通度が一定値以下の文書の集合をＤ’とする。文書集合Ｄ’はその時点で存在するどのクラスターにも属さない公算の大きい文書の集合である。共通共起行列ＴＡ、ＱＡ、文書頻度行列Ｕは最初の繰り返しにおいては全文書Ｄに基づいて計算され、２回目以降の繰り返しにおいては、文書集合Ｄ’に基づいて計算される。一方、クラスターの種となる文書はどの話題の文書が選ばれようと、その話題の中では中心的な文書であることが望ましい。本発明では、Ｄ’の中で最も優勢な話題の文書群において中心的な文書は、Ｄ’との共通度も高いであろうとの仮定のもとに、文書集合Ｄ’中の文書と文書集合Ｄ’との文書共通度を求め、文書共通度の高い文書をクラスターの種の候補として選択する。任意の文書をＰ、その共起行列をＳＰとするとき、文書ＰとＤ’との文書共通度としては例えば以下を用いることができる。

【００３４】

【数１４】

$$com_q(D', P; Q^A) = \frac{\sum_{m=1}^{M}\sum_{n=1}^{M} Q^A{}_{mn} S^P{}_{mn}}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (Q^A{}_{mn})^2}\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{M} (S^P{}_{mn})^2}}$$

．．．（２）

【００３５】

式（２）において行列ＱＡの代わりに行列ＴＡを用いることもできる。また、式（２）において、複数の話題に共通する用語の影響を軽減するために、共起行列、共通共起行列の対角成分は用いないようにしてもよい。

【００３６】

クラスターの種文書の候補は、Ｄ’中の全ての文書に対して式（２）により文書共通度を求め、文書共通度の高い一定個の文書を選択することにより求められる。次にクラスターの種文書の決定について説明する。先ず各候補文書についてＤ’中の全ての文書との類似度を求める。次いで各候補文書について各候補文書との類似度が一定値より大きい文書を各候補文書の近隣文書として求める。近隣文書数が最も大きい文書を候補文書の中から一文書選択することにより、クラスターの種文書が決定される。また、初期クラスターはその種文書の近隣文書で与えられる。

【００３７】

クラスターの成長処理２５では、クラスターの初期クラスターと共通度の高い文書を吸収することによりクラスターを成長させる。図３はそのようなクラスターを成長させる処理のブロック図である。３１は文書頻度行列作成、３２は共通共起行列作成、３３は用語・用語対の特有度算出、３４は文書共通度算出、３５はクラスターメンバー決定、３６は終了条件チェックである。

【００３８】

文書頻度行列作成３１、共通共起行列作成３２では、その時点で一時的に着目クラスターのメンバーとなっている文書集合に対して、図２のブロック２３における文書頻度行列作成処理、共通共起行列作成と同等の処理を行う。３１で求め

られた文書頻度行列をＵにより表す。３２で求められた共通共起行列をＴＡもしくはＱＡにより表す。用語・用語対の特有度算出３３では、各用語・用語対の特有度を決定し、重みを決定する。先ず、前述のように、U^0_{mm}/U_{mm}を用語ｍの文書頻度比として求め、Ｕから求められる最も頻度の高い一定個の用語のうち、文書頻度比の値の小さな一定個を着目クラスターの特有用語として選択する。次いで、これらの用語の文書頻度比の平均を平均文書頻度比ｃ’として求め、用語ｍの特有度v_{mm}、用語対ｍ、ｎの特有度v_{mn}を以下により決定する。

【００３９】

【数１５】

$$v_{mm} = c' / (U^0_{mm} / U_{mm})$$

【００４０】

【数１６】

$$v_{mn} = c' / (U^0_{mn} / U_{mn})$$

【００４１】

あるいは、特有用語に限定することなく、特有用語対と特有用語の両方を用いて平均文書頻度比を求めるようにしてもよい。この場合にはU^0_{mn}/U_{mn}をｍとｎが等しくないときは用語対ｍ、ｎの文書頻度比として、ｍとｎが等しいときは用語ｍの文書頻度比として求め、Ｕから求められる最も頻度の高い一定個の用語、用語対のうち、文書頻度比の値の小さな一定個を着目クラスターの特有用語または用語対として選択する。次いで、これらの用語、用語対の文書頻度比の平均を平均文書頻度比ｃ’として求めるようにする。

【００４２】

用語ｍ、用語対ｍ、ｎの重みをそれぞれz_{mm}、z_{mn}とする。これらは、重み決定関数ｆ（ｘ）を用いて、以下のように決定する。

【００４３】

【数１７】

$$z_{mm} = f(v_{mm})$$

【００４４】

【数１８】

$$z_{mn} = f(v_{mn})$$

【００４５】

ｆ（ｘ）の決め方は任意であるが、ｆ（ｘ）＝ｘ、ｆ（ｘ）＝ｘ^{2}のようにするのがひとつの方法である。あるいは、ｘが一定値よりも大きいときはｆ（ｘ）＝１、そうでないときはｆ（ｘ）＝０としてもよい。

【００４６】

文書共通度算出３４では、全入力文書に対して着目クラスターとの文書共通度を算出する。任意の文書をＰ、その共起行列をＳＰとするとき、文書Ｐの文書共通度は、

【００４７】

【数１９】

$$com_l(D,P;Q^A) = \frac{\sum_{m=1}^{M} z_{mm} Q^A{}_{mm} S^P{}_{mm}}{\sqrt{\sum_{m=1}^{M} z_{mm} (Q^A{}_{mm})^2} \sqrt{\sum_{m=1}^{M} z_{mm} (S^P{}_{mn})^2}}$$

【００４８】

もしくは、

【００４９】

【数２０】

$$com_q(D,P;Q^A) = \frac{\sum_{m=1}^{M} \sum_{n=1}^{M} z_{mn} Q^A{}_{mn} S^P{}_{mn}}{\sqrt{\sum_{m=1}^{M} \sum_{n=1}^{M} z_{mn} (Q^A{}_{mn})^2} \sqrt{\sum_{m=1}^{M} \sum_{n=1}^{M} z_{mn} (S^P{}_{mn})^2}}$$

【００５０】

により求めることができる。上式において行列ＱＡの代わりに行列ＴＡを用いることもできる。

【００５１】

クラスターメンバー決定３５では、ブロック３４で求められた各文書の着目クラスターに対する文書共通度を用いて、一定値以上の文書共通度を有する文書を求め、着目クラスターの一時的なメンバーとする。

【００５２】

終了条件チェック３６では、着目クラスターの成長処理を終了するか否かをチ

ェックする。先ず、初回の繰り返し、即ち初めて３６に到達したときには無条件に３１に戻って処理を繰り返す。２回目以降の繰り返しの時には、上記３５で求められた着目クラスターのメンバーの文書数をカウントし、それが前回の繰り返しのメンバー数と等しくない場合には３１に戻って処理を繰り返す。等しければ、全入力文書と着目クラスターとの文書共通度を保持し、着目クラスターの成長処理を終了する。

【００５３】

図２に戻って説明を続ける。残存文書検出２６では、その時点で存在する全てのクラスターに対する各文書の文書共通度をもとに、どのクラスターに対しても文書共通度が一定値以下の文書を残存文書として抽出する。

【００５４】

終了条件チェック２７では、残存文書数をもとにクラスターの種の検出から成長に到る一連の処理を終了するか否かのチェックを行う。例えば、残存文書数が０、もしくは残存文書数が一定値以下でかつ前回の繰り返し時の残存文書数と等しい場合にはブロック２９に移行するようにする。もし、このような条件が満たされなければ、ブロック２８に移行し、残存文書集合に対してブロック２３と同等な処理を行う。

【００５５】

帰属クラスター決定２９では、各文書が最終的に求められたクラスターのどれに帰属するかを決定する。これは、図３において各文書について求められた各クラスターに対する文書共通度の情報を用い、各文書は文書共通度が最も高くなるクラスターに帰属させることで実行できる。

【００５６】

冗長クラスターの検出・除去３０では、冗長なクラスターが存在するか否かをチェックし、存在する場合には除去する。冗長なクラスターは、ひとつの話題に対して２つ以上のクラスターが求められたときに発生する。そのような時、その話題を記述した文書は２つ以上のクラスターに対して大きな文書共通度を有するようになり、２つ以上のクラスターが重複する格好になる。冗長なクラスター検出のためには、先ず、求められた全てのクラスターと全ての文書との文書共通度を

求め、次いで文書共通度が着目クラスターに対しては一定値よりも大きく、他のどのクラスターに対しても一定値よりも小さくなる文書数を求める。そのような文書数は着目クラスターが他のクラスターと重複しない場合には、その着目クラスターと一定値以上の文書共通度を有する文書数と等しくなる。一方、着目クラスターが他のクラスターと重複する場合には、他のクラスターと重複していない文書数すなわち当該着目クラスターにのみ属する文書数となる。このような文書数を各クラスターの重要度と定義する。重要度は、図４（ａ）の場合「クラスター１」及び「クラスター２」に属する文書数である。重要度は、着目クラスターが他のクラスターと一部重複して存在する場合には、他のクラスターと重複していない文書数となる。すなわち、「クラスター１」に対しては、図４（ｂ）の「ｃ」で示される部分に含まれる文書数を意味する。「クラスター２」に対しては、図４（ｂ）の「ｄ」で示される部分に含まれる文書数を意味する。ひとつのクラスターの重要度が一定値よりも小さい場合は、そのクラスターに属する文書の数がいくら多くとも冗長なクラスターとみなし除去する。そのようなクラスターが複数存在すれば、クラスター重要度が最も小さいクラスターを先ず除去する。その後残されたクラスターについてクラスター重要度の算出を行い、クラスター重要度が最も小さいクラスターを除去する。このような処理を冗長なクラスターが存在しなくなるまで繰り返す。クラスターの削除を行った場合は、除去されたクラスターに属していた文書の帰属クラスターの決定を改めて行う。

【００５７】

【発明の効果】

ここで本発明の効果を説明する為に図２、３の実施例に沿った実験結果を示す。用いたコーパスはＴＤＴ２である。ＴＤＴ２は１９９８年の１月から６月の間の１００個のイベントに関するニュースストーリーの集合であり、６個のニュースソースから採取されている。本報告では同じくＴＤＴ２を用いて行われたＬｉｕらの非階層型のクラスタリング（Ｘ．Ｌｉｕ，Ｙ．Ｇｏｎｇ，Ｗ．Ｘｕ ａｎｄ Ｓ．Ｚｈｕ．Ｄｏｃｕｍｅｎｔ Ｃｌｕｓｔｅｒｉｎｇ ｗｉｔｈ Ｃｌｕｓｔｅｒ Ｒｅｆｉｎｅｍｅｎｔ ａｎｄ Ｍｏｄｅｌ Ｓｅｌｅｃｔｉｏｎ Ｃａｐａｂｉｌｉｔｉｅｓ．Ｉｎ Ｐｒｏｃｅｅｄｉｎｇｓ ｏｆ ｔｈｅ ２５

th Annual International ACM－SIGIR Conference on Research and Development in Information Retrieval, pp. 191－198. Tampere, Finland, August, 2002.）の結果と比較するため、Ｌｉｕらが行ったようにＡＢＣ、ＣＮＮ、ＶＯＡから採取された１５イベントに関するニュースストーリーの集合を実験対象とした。表１にそれらの詳細を示す。

【００５８】

【表１】

イベントID	イベントの内容	文書数			
		ABC	CNN	VOA	Total
01	Asian Economic Crisis	27	90	289	406
02	Monica Lewinsky Case	102	497	96	695
13	1998 Winter Olympic	21	81	108	210
15	Current Conflict with Iraq	77	438	345	860
18	Bombing AL Clinic	9	73	5	87
23	Violence in Algeria	1	1	60	62
32	Sgt. Gene McKinney	6	91	3	100
39	India Parliamentary Election	1	1	29	31
44	National Tobacco Settlement	26	163	17	206
48	Jonesboro shooting	13	73	15	101
70	India, A Nuclear Power?	24	98	129	251
71	Israeli-Palestinian Talks	5	62	48	115
76	Anti-Suharto Violence	13	55	114	182
77	Unabomer	9	66	6	81
86	GM Strike	14	83	24	121

【００５９】

表２に実験に用いられた１４種類のデータセットとそれに対する提案手法とＬｉｕらの手法のクラスタリングの精度を示す。Ｌｉｕらの手法の結果はＬｉｕらの論文より再掲したものである。本発明では、ある文書が属するイベントとその文書が帰属するクラスターの種となった文書のイベントが一致するときクラスタリングの結果は正しいとされる。また、全てのクラスターに対して文書共通度が

０の文書は誤りとする。精度は正しくクラスタリングされた文書数の全文書数に対する比により求める。Ｌｉｕらの方法は、混合ガウス分布モデル（Ｇａｕｓｓｉａｎ　Ｍｉｘｔｕｒｅ　Ｍｏｄｅｌ）に基づき非階層形のクラスタリングを行った後、各クラスターの特有用語を求め、特有用語の投票によって結果を修正している。表２において、テストデータのＡＢＣ－０１－０２－１５とあるのは、ＡＢＣより採取されたイベントＩＤが０１、０２、１５に属する文書であることを意味している。表２から、精度の高いデータセットの数は、Ｌｉｕらの方法よりも本発明の方が多く、本発明が優ることが分かる。

【００６０】

【表２】

番号	データセット	Liuらの方法	本発明
1	ABC-01-02-15	1.0000	0.9806
2	ABC-02-15-44	0.9902	0.9805
3	ABC-01-13-44-70	1.0000	1.0000
4	ABC-01-44-48-70	1.0000	1.0000
5	CNN-01-02-15	0.9756	0.9932
6	CNN-02-15-44	0.9964	0.9964
7	VOA-01-02-15	0.9896	0.9986
8	VOA-01-13-76	0.9583	0.8943
9	VOA-01-23-70-76	0.9453	0.9206
10	VOA-12-39-48-71	0.9898	1.0000
11	VOA-44-48-70-71-76-77-86	0.8527	1.0000
12	ABC+CNN-01-13-18-32-48-70-71-77-86	0.9704	0.9917
13	CNN+VOA-01-13-48-70-71-76-77-86	0.9262	0.9500
14	ABC+CNN+VOA-44-48-70-71-76-77-86	0.9938	1.0000

【００６１】

また、求められるクラスターの個数も表２のデータに対して全て正しく求められている。

また、Ｌｉｕらの論文で挙げられている１２種類のデータに対してもクラスター数は正しく求められている。一方、Ｌｉｕらの方法では１２種類のうち３種類のデータに対して正しく求められていない。表３にＬｉｕらの方法及び本発明の結果を示す。

【表３】

テストデータ	正しいクラスタ数	Liu等のテスト結果	本願発明のテスト結果
ABC−01−03	2	2	2
ABC−01−02−15	3	3	3
ABC−02−48−70	3	2	3
ABC−44−70−01−13	4	4	4
ABC−44−48−70−76	4	4	4
CNN−01−02−15	3	4	3
CNN−01−02−13−15−18	5	5	5
CNN−44−48−70−71−76−77	6	5	6
VOA−01−02−15	3	3	3
VOA−01−13−76	3	3	3
VOA−01−23−70−76	4	4	4
VOA−12−39−48−71	4	4	4

【００６２】

このように本発明によれば、入力文書集合から正しい数のクラスターを抽出すること、及び各文書が帰属すべきクラスターを精度よく決定することができ、ユーザの情報獲得の効率性が高められる。

【図面の簡単な説明】

【図１】

本発明の概略を示すブロック図である。

【図２】

文書集合が入力された段階からクラスターと各文書が帰属するクラスターが決定されるまでの手順を示す図である。

【図３】

一つのクラスターについて初期クラスターの段階からの成長の手順を示す図である。

【図４】

冗長なクラスターを削除する為のクラスターの重要度を説明する図である。

【符号の説明】

１１０：文書入力ブロック

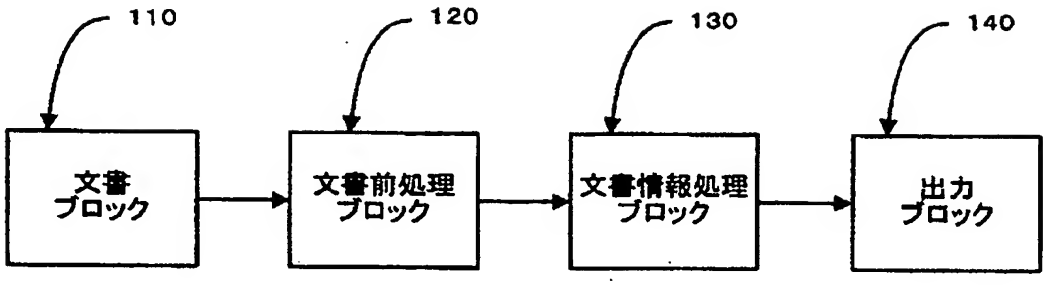
１２０：文書前処理ブロック

１３０：文書情報処理ブロック

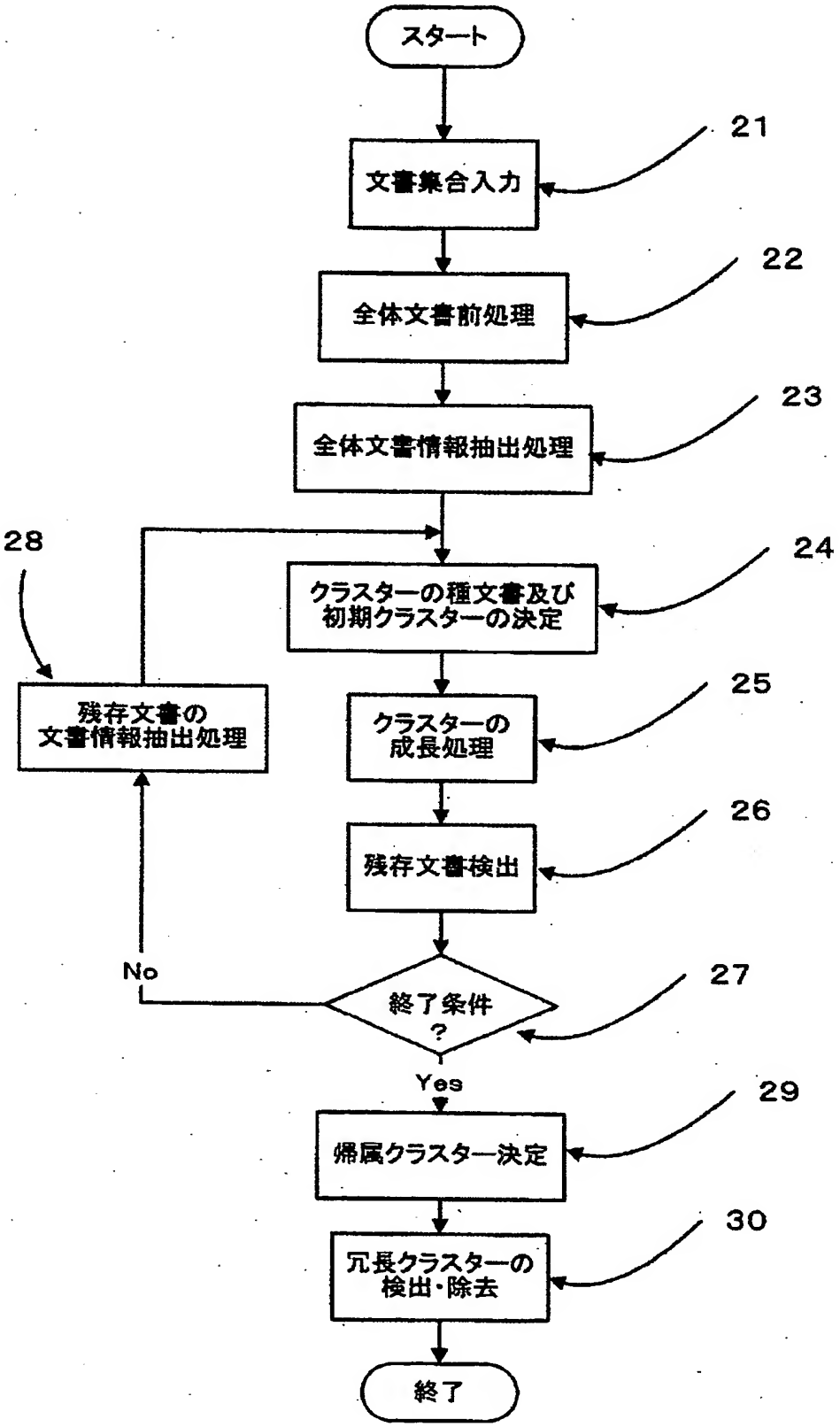
１４０：出力ブロック

【書類名】　図面

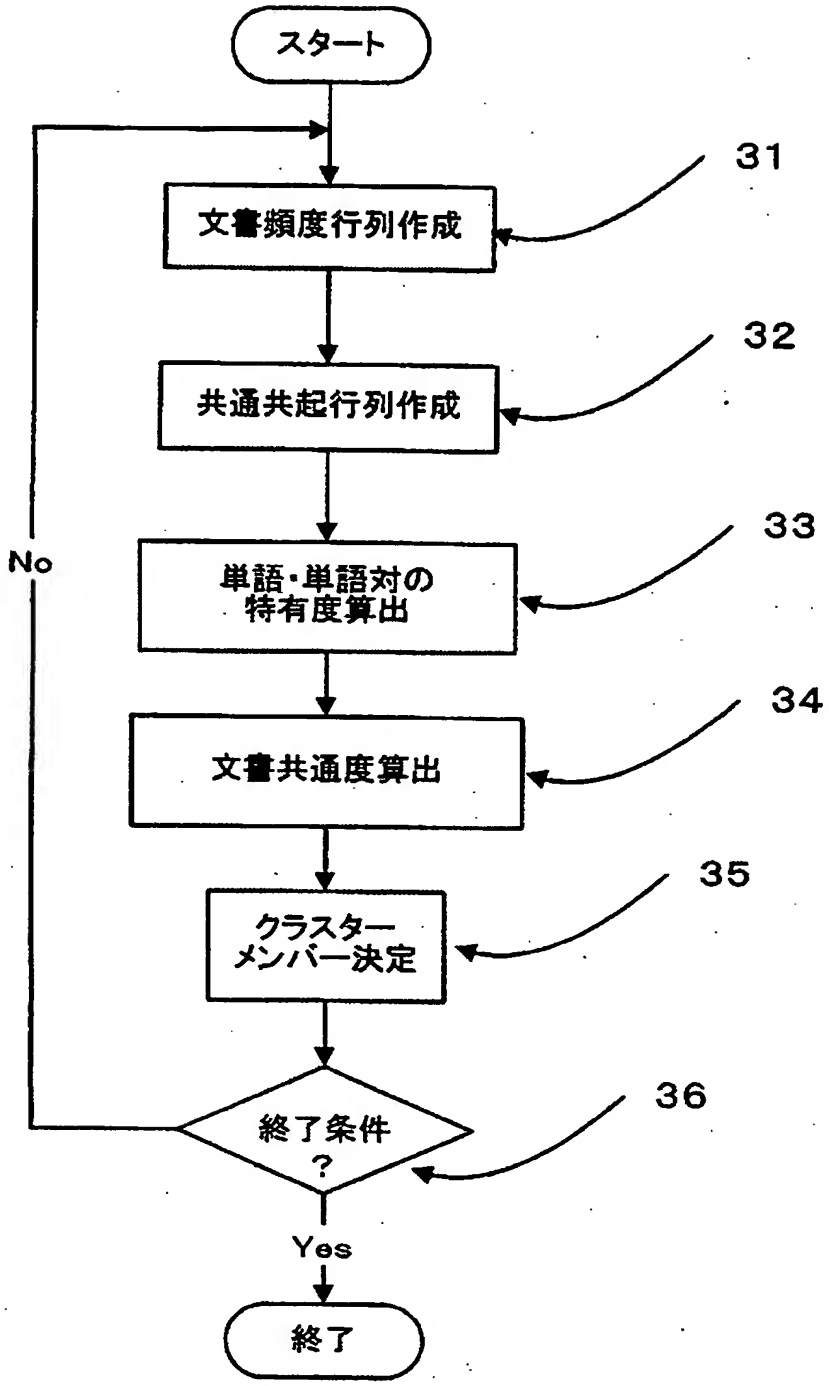
【図１】



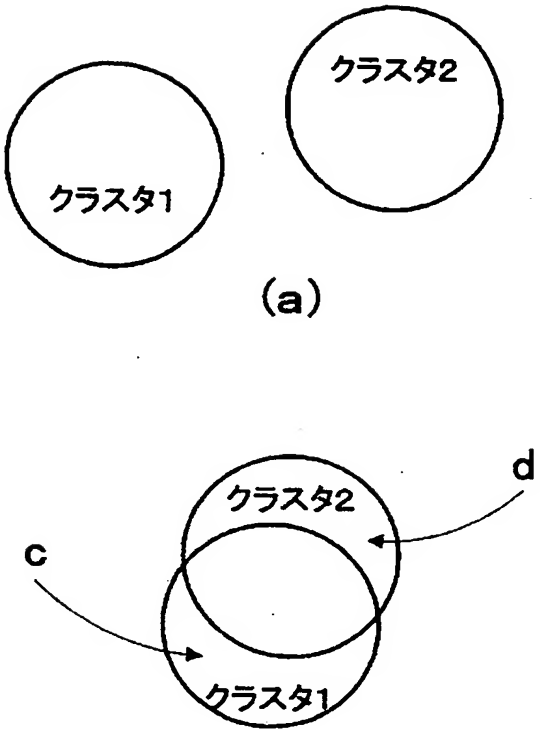
【図２】


スタート
21
文書集合入力
22
全体文書前処理
23
全体文書情報抽出処理
28
24
クラスターの種文書及び
初期クラスターの決定
残存文書の
文書情報抽出処理
25
クラスターの
成長処理
26
残存文書検出
No
27
終了条件
?
Yes
29
帰属クラスター決定
30
冗長クラスターの
検出・除去
終了

【図３】


スタート
31
文書頻度行列作成
32
共通共起行列作成
33
単語・単語対の
特有度算出
34
文書共通度算出
35
クラスター
メンバー決定
36
終了条件
？
No
Yes
終了

【図４】



【書類名】　要約書

【課題】

文書のクラスタリングにおいて、正しい数のクラスターを求めること、及び各文書の帰属するクラスターを精度よく求めることは完全には解決されていない問題であった。

【解決手段】

文書クラスタリングでは同じ話題を述べた文書がグループ化されるので、同じクラスターに属する文書群には何らかの共通性があるはずである。また、各話題には話題特有の用語や用語対が存在する。本発明ではこれらの点に着目し、各文書の着目クラスターへの近さを求めるときに、着目クラスターに特有でない用語や用語対の影響を排除しつつ着目クラスターの共通情報を用いるようにした。

【選択図】図１

特願２００３－１０５８６７

出　願　人　履　歴　情　報

識別番号　　[３９８０３８５８０]

１．変更年月日　　１９９８年　５月１９日
[変更理由]　　新規登録
住　所　　アメリカ合衆国カリフォルニア州パロアルト　ハノーバー・ストリート　３０００
氏　名　　ヒューレット・パッカード・カンパニー